Creating Three-Dimensional Thermal Maps

Mathew Price Cogency cc Cape Town Email: mathew@cogency.co.za Jeremy Green CSIR Centre for Mining Innovation Johannesburg Email: jgreen@csir.co.za John Dickens CSIR Centre for Mining Innovation Johannesburg Email: jdickens@csir.co.za

Abstract—A method for generating 3D maps of mines with thermal imaging texture from a robotic platform is presented. The objective is to use these models to generate risk maps that can be used to assess mine safety. Analysis will take place offline, but the data collection platform will be autonomous.

Our registration method is based on aligning 3D descriptors that are extracted from range images and uses ICP for refinement. The descriptors encode the distribution of radial distances in the vicinity of keypoints, and enable alignment of non-sequential scans that are visibly different. Thermal texture is added to the registered point cloud using a once-off calibration between the 3D and thermal cameras.

Two alternate visualisation schemes are discussed, and results are shown for a real mine stope.

I. INTRODUCTION

The CSIR is pursuing the development of an Underground Autonomous Mine Safety Platform (UAMSP) [1] that will provide insight into aspects of mine safety. In particular, the vehicle will collect thermal and electronic sounding data that can be used to assess hanging wall stability [2], thereby indicating the risk of potential rock falls — a significant cause of injury and death in mines.

One of the challenges involves building a 3D map of the mine stope and overlaying thermal measurements in order to generate a risk map. A framework for registering 3D and thermal data in order to accomplish this task is presented.

A. Background

Previous studies in South African hard rock mines [3], [4], [5] have identified a correlation between loose rock and its surface temperature gradient. Blasting often results in the formation of cracks that separate portions of rock which interrupts the heat flow and causes preferential cooling in a localised area. The effect can be detected by long-wave infrared cameras (thermal imaging) and used to identify candidate areas that are unsafe. Electronic sounding, which requires contact with the surface, can then target these candidate areas to provide more detailed analyses.

The fact that thermal differences can be caused by a number of phenomena, such as surface angle and material composition, motivated a multi-sensor approach. The incorporation of both 3D and thermal imaging cameras enables thermal risk maps to correctly account for surface geometry and provides a mechanism for mapping the environment. (An important component for assisting robot navigation and facilitating visualisation.)

B. Approach

Our process for generating a thermal map is as follows: Given a sequence of range images (where pixels encode depth), each scan (image) is registered to a common reference frame. Once aligned, the points are projected into the corresponding thermal images, thereby associating a temperature measurement with each 3D point. Finally, the map is textured with the temperatures and rendered for visualisation.

Registration is tackled in two steps: 3D descriptors are used to establish an initial solution, and ICP (iterative closest point) [6] is used for refinement. In addition, we follow a similar approach to [7] where keyframes are used to limit model complexity. Using descriptors allows detection and alignment of arbitrary partially matching scans even when there is a substantial difference in relative camera pose. This method lends itself well to other issues such as loop-closure since constraints can be established when portions of the map are revisited.

The projective mapping between the 3D and thermal cameras is determined using a calibration step that is applied onceoff. This is possible because the cameras are rigidly mounted together.

C. Data Acquisition

The availability of low-cost 3D cameras, such as the Microsoft Kinect, has generated a lot of recent interest, because the acquisition of 3D data is now more accessible to the average user. The Kinect, which is designed for real-time human motion capture for games, comprises a colour camera and an infrared camera-projector pair that is able to produce calibrated range data. In our experiments, we have used the Kinect depth camera to acquire a number of 3D underground scans, and found that it works well in the unlit mining environment. Our test rig (Figure 1) also includes a SR4000 time-of-flight (TOF) camera and a FLIR A300 thermal imaging camera. (The inclusion of two 3D cameras is experimental, and ultimately only one will be retained.)

II. SCAN REGISTRATION

Estimating the relative alignment between scans (essentially, the camera's ego-motion) is a crucial component for mapping. Imagine that each scan is a piece of a 3D puzzle. To create the map we need to orient each piece (3D rotation) and place it (3D translation) so that it correctly connects with neighbouring



Fig. 1. Capture Setup: Microsoft Kinect (top); Mesa SR4000 (left); FLIR A300 (right).

pieces. It is not difficult to see that misaligning just one piece will result in an inaccurate map. Moreover, if we base the poses of new pieces on the current map, one previous misalignment will cause a number of subsequent errors. Therefore, in order to create an accurate map we need: a method to choose the pose for each scan (alignment); and a way to ensure that the collective set of poses is globally consistent. The latter is often referred to as SLAM (Simultaneous Localisation and Mapping), and is an active area of research.

A. ICP Alignment

An efficient way of aligning two point sets is to use iterative closest point (ICP) [6]. The vanilla method is based on a two-step procedure in which points are first associated with their nearest neighbours in the opposite set, and a Euclidean transform is computed that minimises the error between the pairs. The procedure is repeated until the transformation converges, but this is only guaranteed in the noiseless case.

In the case of robot mapping, an additional issue is the motion of the scanner which produces partially matching scans. To cope, a distance threshold is used to omit pairings with large separation, so that only common structure drives the alignment. However, the trade-off is that ICP will only succeed when the relative difference between scans is small. Thus, a standard approach is to use ICP to align successive scans that are temporally close. (I.e. It is assumed that the robot moves slow enough such that consecutive scans are similar.)

An ICP-only approach has the advantage of speed (with the use of various optimisations [8]), but suffers from a degenerate global solution due to the build up of accumulated errors. Another problem is that it is difficult to flag bad alignments, since it is possible to obtain a low error for partially matching scans even when the transformation is incorrect. This makes the process sensitive to outliers. For this reason, we use a hybrid approach where an initial transformation is computed with 3D descriptors, and ICP is used for refinement.

B. Visual SLAM

Since our data currently comprises range and thermal images, it is not possible to use odometry and inertial measurements to compensate for ICP errors. Therefore, an alternative method for ensuring robust registration was sought. Using invariant feature methods [9] is a popular SLAM technique for robots equipped with colour cameras. Efficient methods such as SIFT [10] and SURF [11] enable robust detection and matching of interest points in images. By targeting point-like features, these *keypoints* can be detected across substantial differences in viewing angle, and use the local pixel neighbourhood to create a fingerprint for matching. More recently, this technique has been extended to use both colour and depth cameras [7]. This enables 3D localisation of the visual features and less constraints on the pose estimation process.

While visible-light features are not ideal for mines, due to illumination issues, we have had success in registering thermal images with SURF. However, the existence of sufficient thermal features is very unpredictable, leading to the conclusion that this can only be used as an assistive measure.

C. 3D Descriptors

An alternative to blindly aligning point clouds with ICP is to find characteristic structures in the 3D data, and use this to establish a canonical reference frame. This is what 3D descriptors are designed to do, and there are many uses: object recognition, 3D content retrieval, and alignment, to name a few. Unsurprisingly, there are many commonalities with 2D invariant feature methods.

Good descriptors should be: fast to compute; easy to compare; invariant under rotation and scale; and should provide good discrimination. (Achieving fast computation is often difficult for 3D methods.) For 3D data, scale invariance is implicit since most 3D devices are calibrated.

Two popular examples of 3D descriptors are 3D Shape Contexts [12] and Spin Images [13]. Both of these methods are based on generating rotation invariant histograms around interest points that can be easily compared. 3D Shape Contexts use a spherical neighbourhood with logarithmic bin spacing in the radial direction, which allows descriptors to have a large sphere of influence while ensuring that closer features are given more weight. Spin Images are similar, but use a linear cylindrical coordinate frame. However, the methods are computationally intensive and rotational invariance is achieved by aligning the local reference frame to the surface normal, which results in the descriptor being very sensitive to small changes in the estimated normal. Since these methods are often applied to high-resolution scans or synthetic data this is not problematic. This is not the case for noisy scanners, such as the Kinect or SR4000 time-of-flight camera, used here.

Aligning the descriptor's reference frame to the local surface normal removes 2 degrees of freedom (i.e. planar alignment with unknown azimuth). Since it is difficult to uniquely specify the remaining rotation about the normal (with respect to the descriptor), a standard tactic is to do a 1D search to determine the best correspondence during matching. We dislike this for computational reasons, but are especially wary of the reliance on the local surface normal which has proved to be noisy in our data. This seems to be more problematic for the timeof-flight camera data than for the Kinect, but both systems perform badly near sharp edges.

While the aforementioned methods are good descriptors, they do not specify a method for generating interest points. This is important since the penalty of higher computation needs to be mitigated by computing less features — at least when applied to online scan registration. Applying SIFT or SURF directly to range images does not work since pixels encode relative distances to the camera centre which changes depending on where the camera is located. Therefore, some reformulation is necessary. In [14], Lo and Siebert use the shape index (the ratio of principal curvatures) to locate keypoints in scale-space. For descriptors, they use a histogram of local surface shapes, e.g. saddle, trough, ridge, dome. A more direct extension to 2D SIFT is presented by Gibbins [15] where he computes the 3D Hessian and uses a variation of Spin Images as descriptors.

Once again, both these extensions are heavily dependent on surface curvature. (We are especially sceptical of the second approach where results were only conducted on simulated data.) Furthermore, since 2D SURF already pushes processing requirements, we felt it unlikely that the 3D extensions are suited for online usage. Since we only require the ability to match scans that are likely to differ by small amount, we opted for a simpler descriptor.

III. REGISTRATION METHOD

Drawing on insight from [7] and [16], our registration method uses 3D descriptors to establish an initial alignment between scans, after which ICP is used to refine the solution. Other ideas that were adopted include: downsampling the Kinect range images by 8 (Interestingly, this still provides enough detail for ICP.); and using keyframing to reduce the complexity of the model.

Our variation of keyframing differs slightly from suggested approaches in that we also use it as a means to limit accumulated error. (This is because not all our data sets comprise closed loops.) New keyframes are initialised whenever the number of descriptor matches falls below a threshold. Otherwise, descriptors are used to establish a transform to the last keyframe and then ICP is applied for optimisation. This approach, illustrated in Figure 2, also allows us to skip frames to increase processing speed, since the full sequence of range images is often not required. The set of keyframes thus provides a light-weight representation.

The following is a breakdown of the registration procedure:

- Assign first scan as keyframe F_k and set initial keyframe pose to M_k = I
- 2) Apply descriptor alignment (DA) between F_i and F_k to obtain relative pose M_i
- 3) If DA succeeds:
 - a) Use M_i as starting point for ICP and obtain refined pose M_{icp} (relative to initial pose)
 - b) Store global pose $M_k M_{icp} M_i$ for F_i

4) Else:

a) Use above method to register F_i to F_{i-1} and set as new keyframe



Fig. 2. Keyframing: Sequential frames are matched against the last keyframe. New keyframes are initialised when the number of descriptor matches falls below a threshold. The sequence of keyframes forms an efficient, light-weight representation of the model.

b) Terminate on end of sequence or failure to create a new keyframe

(where M_x represents a 4x4 rigid body transform, and global poses are specified in the reference frame of the first keyframe.)

Descriptor alignment is based on a novel 3D descriptor, namely the distance signature (dSig), which is described shortly.

A. Keypoint Selection

Prior to generating descriptors, we must first choose suitable keypoints.

Methods such as SURF and SIFT construct a scale-space using the input images and choose points that can be consistently detected. We found that this process is relatively time consuming (OpenCV SURF takes around 1 second on a 640x480 image), since full resolution images are required for best subpixel localisation. Even then, scanner noise causes poor repeatability of keypoints. Similar issues are encountered with the 3D extension methods discussed earlier, but we continue to explore options.

For now, we leverage the richness of the 3D descriptors and adopt a uniform sampling strategy. Uniform sampling effectively ignores the localisation problem, but suffers from reduced precision since keypoints do not consistently target features at a subpixel level. This also affects repeatability since there are no guarantees that the same features are sampled in subsequent frames. However, the reduced processing time makes up for this by enabling a dense field of descriptors to be captured, ensuring that most of the scene is characterised. Our current naive implementation takes approximately 70ms to extract 1500 features from a downsampled frame.

B. The Distance Signature (dSig)

A signature [17] is an efficient representation of a distribution in which bins are selected in order to give more descriptive weight to denser parts of the distribution. This reduces quantisation error that plagues histogram methods, and provides good descriptive power. In the 1D case, a signature can be thought of as a histogram with bins of equal count instead of equal width. It can be easily computed by dividing a sorted list of data among a number of bins and computing the mean of each bin. The result is a sampled version of the cumulative distribution function which has a useful property: The area between two 1D signatures (i.e. the L_1 norm) is equivalent to the Earth Mover's Distance (EMD). EMD has been shown to provide good performance as a similarity measure, and has been used extensively in applications such as image-based retrieval [17].

Consequentially, our descriptors are signatures that encode the distribution of distances within a specified radius of a keypoint. This offers: fast computation, invariance under rotation, and resistance against sensor noise. However, the lack of angular information does reduce descriptive power.

For efficiency, our implementation operates directly on the downsampled range images. A processing window is selected around each keypoint where the width of the window (in pixels) is selected so that its 3D width corresponds with the radius of interest:

$$w_{\rm 2D} = \frac{f}{d} w_{\rm 3D},\tag{1}$$

where f is the effective focal length, d is the depth from the camera, and w is the width of the window. This provides a fast means of obtaining candidate neighbours within the sphere-of-influence directly from the image. Distances within the radial threshold are used to construct the signature, as illustrated in Figure 3. Generally, more bins results in a higher descriptor



Fig. 3. Generating Distance Signatures. Top-left: Range image with uniform sampling grid and search window for one of the keypoints. Top-right: Signatures generated for each keypoint. Bottom: Illustration of how the 2D search window corresponds to the keypoint's sphere-of-interest. A radius of 80cm is used.

resolution. Currently, we use signatures with 64 bins.

C. Pose Estimation

Given a set of dSig features for two range images, we use RANSAC (Random Sample Consensus) to determine their relative pose. First, matching features are collected using the EMD between signatures. A greedy approach is employed whereby matches are assigned on a lowest-EMD basis.

Sets of 3 matching keypoints are repeatedly sampled and a Euclidean transform that aligns them is computed. The candidate solution is used to transform the remaining keypoints and the number of inliers (matching keypoints that are within a specified tolerance) is counted. The solution that results in the maximum number of inliers is selected. We do not include a final optimisation step, but instead apply ICP on the original downsampled point clouds to refine the RANSAC pose. This compensates for the poor keypoint localisation.

An example of estimating the pose between two nonsequential range images (separated by 16 degrees) is shown in Figure 4. RANSAC is applied to both sets of signatures and



Fig. 4. Pose alignment using dSig and RANSAC. Top images: Non-sequential range images with a relative viewing angle of 16 degrees. Green markers show RANSAC inliers. Bottom-left: Synthetic range image generated from the registered model (i.e. 3D points from both images). Bottom-right: 3D view of registered model and showing camera poses.

inliers (25) are shown in green. The lower image shows the quality of alignment by synthesising a range image using the combined model, while the 3D plot shows the aligned point clouds.

IV. THERMAL TEXTURE

The main difficulty in associating thermal data with the 3D model is determining the relative transform between the cameras. This is straightforward for stereo colour cameras, but 3D and thermal cameras do not share a common spectrum. Therefore, our approach involves tracking a known object whose structure can be found in range images, but also has a thermal signature. Our calibration object consists of a ball mounted on a stick, which can be easily segmented using circle

detection. Rubbing the ball (or dipping it in heated water) is sufficient to make the ball visible to the thermal imaging camera.

Once a calibration has been determined, points from the 3D model are projected into each thermal image and their temperatures are computed using bilinear interpolation. To ensure that occluded surfaces are correctly textured, we enforce visibility checking by only retaining temperatures that correspond to the thermal image where the camera is closest.

V. VISUALISATION

Visualisation is an important output of the system because offline analysis will be conducted. We use two methods depending on the application.

A. Viewpoint Mesh

For instances where the area of interest is limited and higher detail is required we use viewpoint mesh projection. This involves projecting the registered model to a planar view and rendering a synthetic range image. Pixel connectivity in the artificial view is used to specify a triangular mesh. The use of a projected view means that meshes can only be generated for a specified view of the model. A more general model can be generated using surfels.

B. Surfels

A problem with manifold meshes is that adjustments are complicated. This is because the connectivity must be manipulated while ensuring that the topology is not compromised. (E.g. faces should not intersect.) Point clouds, on the other hand, are much easier to update, but are difficult to visualise since everything becomes jumbled (due to the interference of background points).

Surfels [18], or surface elements, offer a compromise by combining properties from both representations. The idea is to represent the model as a point cloud, but with extra attributes for each point: surface normal, radius, and colour. This produces a sparse representation that can be rendered in a mesh-like manner without the hassle of interconnectivity — essentially, polygon-soup. Adopting strategies from [19] and [7], we use circular surfels. For efficiency, each circle is approximated by a hexagon that is constructed from four triangles as shown in Figure 5.



Fig. 5. Left: A single circular surfel composed of 4 triangles. Right: A hemisphere represented by surfels derived from points and normals.

Surfels are ideal for building incremental models since each scan can be applied in an iterative manner. The model is adapted using a three step formula: surfel update, surfel addition, and surfel removal. For brevity we refer the interested reader to the cited papers for further details.

VI. RESULTS

We have applied our method to a number of real data sets of both indoor and underground environments. (Videos of some results are available on YouTube [20].) Qualitative assessment is used since ground truth is not available and gross errors can be easily identified. However, it is planned to use recently acquired laser-scanner data for validation in future.

Figure 6 shows thermal map outputs using viewpoint projection for a data set captured inside a mine stope at the Bafokeng Rasimone Platinum Mine (BRPM). The data was captured approximately 12 hours after blasting at a depth of 250 metres. The test rig is mounted on a tripod and manipulated by hand,



Fig. 6. Thermal map of a mine stope. Top: Depth map of rendered view; Middle: Combined thermal texture for model (radiometric scale); Bottom: Textured 3D mesh. Registration uses downsampled points, but the output is generated using the full resolution range data.

resulting in a pan-tilt motion. After registration, the subset of keyframes is used to generate a combined model and the thermal texture is computed. In the figure, the top image shows a depth map (synthesised from the registered model) corresponding to the virtual view; the merged 2D thermal image is shown in the middle; and the viewpoint mesh of the thermal map is shown at the bottom. Misalignments are caused by registration error and timestamp differences between the 3D and thermal images. The latter is because the FLIR A300 is a 3Hz camera (which is more cost-effective) and the nearest Kinect frame (30Hz) is used when computing the texture. This currently makes it difficult to use the thermal images to refine the registration solution, but this is being investigated.

For comparison, Figure 7 shows the thermal map generated using surfels. While the mesh version is better for detail, the surfel model is light-weight and better for incremental modelling of large areas.



Fig. 7. Thermal map rendered using surfels. Top: Surfels are coloured by depth to the origin; Bottom: Surfels are coloured using thermal measurements (radiometric).

VII. CONCLUSION

A processing framework has been presented for generating 3D thermal maps of mines using range and thermal images.

Our method uses a novel 3D descriptor, the distance signature, to align scans and ICP is used for refinement. In addition, keyframing is used to slow the accumulation of incremental errors and construct a light-weight model.

Results have been shown where a 3D map with thermal texture is generated for a real mine stope. Small variances in pose estimates that reduce model precision do occur, and timestamp differences between the cameras also contributes to alignment error. However, for the most part reasonable registration is achieved using only the range images.

Future research will explore further optimisations and integration with secondary sensors such as inertial measurement devices.

ACKNOWLEDGEMENTS

We thank the CSIR SRP for continuing to support this work and the brave team of gatherers who ventured below to bring us data.

References

- J. J. Green, P. Bosscha, L. Candy, K. Hlophe, S. Coetzee, and S. Brink, "Can a robot improve mine safety?" in 25th International Conference on CAD/CAM, Robotics and Factories of the Future (CARsFOF), 2010. [Online]. Available: http://hdl.handle.net/10204/5022
- [2] D. Vogt, V. Z. Brink, S. Brink, M. Price, and B. Kagezi, "New technology for improving entry examination, thereby managing the rockfall risk in South African gold ad platinum mines," in *CSIR 3rd Biennial Conference*, 2010. [Online]. Available: http: //researchspace.csir.co.za/dspace/bitstream/10204/4255/1/Vogt_2010.pdf
- [3] V. A. Kononov, "Pre-feasibility investigation of infrared thermography for the identification of loose hangingwall and impending falls of ground," CSIR Division of Mining Technology, Tech. Rep., 2000. [Online]. Available: http://www.mhsc.org.za
- [4] —, "Infrared thermography of loose hangingwalls," CSIR Division of Mining Technology, Tech. Rep., 2002. [Online]. Available: http://www.mhsc.org.za
- [5] D. Vogt, V. Z. Brink, and S. Schutte, "New technology for real-time in-stope safety management," in *Hard Rock Safe Safety Conference*, 2009. [Online]. Available: http://researchspace.csir.co.za/dspace/handle/ 10204/3680
- [6] P. J. Besl and H. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=121791
- [7] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D Mapping : Using Depth Cameras for Dense 3D Modeling of Indoor Environments," Work, vol. 1, no. c, pp. 9–10, 2010. [Online]. Available: http://www.seattle.intel-research.net/RGBD/ RGBD-RSS2010/papers/henry-RGBD10-RGBD-mapping.pdf
- [8] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," Proceedings Third International Conference on 3D Digital Imaging and Modeling, vol. pages, pp. 145–152, 2001. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=924423
- [9] R. Munguia and A. Grau, "Monocular SLAM for Visual Odometry," in 2007 IEEE International Symposium on Intelligent Signal Processing. IEEE, 2007, pp. 1–6.
- [10] D. Lowe, "Distinctive image features from scale-invariant keypoints," 2003.
- [11] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in 9th European Conference on Computer Vision, Graz Austria, May 2006.
- [12] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik, "Recognizing Objects in Range Data Using Regional Point Descriptors," *Current*, vol. 1, pp. 224–237, 2004.
- [13] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3D scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 433–449, 1999.
- [14] T. Lo and J. P. Siebert, "Local feature extraction and matching on range images: 2.5D SIFT," *Computer Vision and Image Understanding*, vol. 113, no. 12, pp. 1235–1250, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2009.06.005
- [15] D. Gibbins, "3D Target Recognition Using 3-Dimensional SIFT or Curvature Key-points and Local SPIN Descriptors," in *Defence Applications of Signal Processing 2009 (DASP'09)*, Kauai (Hawaii), 2009.
- [16] T.-W. R. Lo and J. Siebert, "SIFT keypoint descriptors for range image analysis," *Methodology*, vol. 2008, no. 3, pp. 1–17, 2008.
- [17] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.
- [18] H. Pfister, M. Zwicker, J. Van Baar, and M. Gross, "Surfels: Surface elements as rendering primitives," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, ser. Computer Graphics Proceedings, Annual Conference Series, K. Akeley, Ed., vol. pp. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 335–342.
- [19] T. Weise, T. Wismer, B. Leibe, and L. Van Gool, "In-hand scanning with online loop closure," 2009 IEEE 12th International Conference on Computer Vision Workshops ICCV Workshops, pp. 1630–1637, 2009.
- [20] M. Price, "Published Videos." [Online]. Available: http://www.youtube. com/matprice78